Combinational Meta-paths Mining for Correlation Relationship Evaluation in Bibliographic Networks

Qinchen Wu, Peng Wu, and Li Pan(✉)

National Engineering Laboratory for Information Content Analysis Technology, School of Electric Information and Electrical Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Rd, Shanghai, China
panli@sjtu.edu.cn

Abstract. Correlation relationships between objects are pervasive in heterogeneous information networks such as bibliographic networks, which made it possible to evaluate proximity between nodes from different perspectives. To explain these semantically rich correlations, meta-paths formed by interconnected node types and edge types have been widely used. This means, using meta-paths and their combinations we can explicitly evaluate relationships between nodes, and thus made it possible to search for proximate nodes according to specific correlations they carried. In this paper, we propose a combinational meta-paths mining algorithm to evaluate correlation relationships between nodes in bibliographic networks. Experiments with bibliographic networks have proved its effectiveness with respect to prior knowledge based results.

Keywords: Heterogeneous information networks · Meta-paths
Correlation relationships · Node-pairs proximity

1 Introduction

Heterogeneous Information Networks (HIN) are graph structures containing multiple typed nodes and edges, where diversities in node and edge types make them carry rich semantics [1, 2]. To explain these semantics, meta-paths that formed by sequence of node and edge types have been widely used [3–5]. Due to diversity of correlations between nodes in HIN, it is preferred to explain them with combinational meta-paths to ensure their semantical completeness. For example, in bibliographic networks that contain node types Paper (P), conference (C), Author (A) and Topic (T), we can describe correlations carried by example pair (Jiawei Han, Philip S. Yu) using combinational meta-path <A-P-A, A-P-T-P-A, A-P-C-P-A> to explain their correlations. These meta-paths have been widely used in recommendation system, information retrieval, link prediction [6–9] in HINs.

Combinational meta-paths provide us solution to describe correlations between nodes in HINs, but we can't greedily include infinite paths. To tackle this problem, we propose combinational meta-path mining (CMPM) algorithm to describe inter-nodes

I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 660–669, 2018.
https://doi.org/10.1007/978-3-030-00916-8_60

correlations in HINs. This algorithm only requires single example pair as input, and it begins with top-k shortest path searching algorithm to find candidate meta-path that connects source and target nodes. Then, weighting the meta-paths with path instances distribution information, and refining candidate meta-paths by maximizing their correlation semantics. Extensive experiments on two bibliographic networks (ACM and DBLP) have proved its effectiveness. Moreover, CMPM detects new meta-paths that are not able to be given by domain knowledge.

The rest of the paper is organized as follows. Section 2 reviews existing methods in meta-paths mining. Section 3 explains basic concepts and definitions in this paper. Then, implementation details of CMPA algorithm are given in Sect. 4. Section 5 evaluates CMPM's effectiveness by detecting proximate nodes. Finally, we conclude this paper in Sect. 6.

2 Related Work

Meta-path has been widely used in HIN related studies to explain their rich semantics, but dependences on prior knowledge have limited their usages in practical applications. Early research by Yizhou Sun et al. [1] summarize meta-paths mining methods into domain knowledge based methods [9, 10], exhaustive trail-selection methods and statistical learning methods [11, 12]. The first category can give concise meta-paths by domain experts when correlation relationships and HINs are simple enough, but they are fragile to outside interference. The trail-selection methods require exhaustive greedy search for meaningful paths, such operation is time consuming when the given heterogeneous information network is large. The third learning algorithm based category relies more on high quality training data. Specific algorithm such as Path Ranking [11] aims at using pre-provided single path as input to obtain optimal path weighting coefficients. AMIE [13] tries to mining meta-paths based on association rules in data, but suffers from global solution. FSPG [12] iteratively selects currently the most relevant meta-path by using step selection but requires highly quality positive/negative example pairs.

Traditional meta-paths mining algorithms as mentioned previously heavily relies on prior knowledge, which limited their usage in large scale heterogeneous information networks. Other automatic path mining algorithms such as Path Ranking, AMIE are either constrained by uncertain hyper-parameter or high computational expenses. In facing with these facts, our CMPM algorithm only takes single example pair as input. Except for this, CMPM also avoids greedy path searching with YenKSP.

3 Problem Definition

In this section, we introduce basic concepts and definitions in combinational meta-paths mining for correlation relationships description.

Heterogeneous Information Network (HIN) is the core concept in this paper. It contains multiple type of nodes and edges [1]. Bibliographic network is typical HIN.

Definition 1. Heterogeneous Information Network is directed graph structure $G = (V, E, \Phi, \Psi)$ with nodes from V interconnected with each other by edge from E. Functions $\Phi : v \rightarrow A$ and $\Psi : E \rightarrow R$ in the network specifies mapping relationships from node $v \in V$ to node type $\Phi(v) \in A$ and from edge $e \in E$ to edge type $\Psi(e) \in R$.

Compared with traditional homogeneous network, HINs generally carry richer semantics because of $|A| > 1$ or $|R| > 1$. To describe the rich semantics, we use meta-paths [1, 3, 15].

Definition 2. Meta-Path carries correlation semantics between interconnected nodes in HINs. Meta-path is defined as $\Pi^{1,2,\ldots,l} = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_{l-1}} A_l$, where l is the path length, A_i and R_i with $i \in [1, l]$ are node type and edge type respectively.

Definition 3. Combinational meta-paths $F_{(s,t)} = \{(\Pi_i, w_i), 1 \leq i \leq P\}$ carry correlation semantics within the sample, where $\Pi_i, 1 \leq i \leq P$ are P single meta-path, w_i are weights for Πi as a measure of importance, which subject to $\sum_{i=1}^{P} w_i = 1$ and is solved according to path instances distributions.

Using $F_{(s,t)}$ given by (s, t), the correlation semantics intensity vector associated with it is $f_{(s,t)} = \{\omega_i \sigma((x, y)|\Pi_i), 1 \leq i \leq P\}$, where $\sigma((s, t)|\Pi_i)$ is semantic intensity score measured with nodes similarity function. One fact should be strengthened here; we quantify semantics intensity as similarity score because classical node similarity measures such as PathSim [1], HeteSim [10], Path Constrained Random Walk [11] are all meta-paths based, and node-similarity relationship is exception of more general concept of nodes correlations. To find proximate node pair (x, y) that sharing similar correlations as (s, t), we calculate $f_{(x,y)} = \{w_i \sigma(x, y)|\Pi_i, 1 \leq i \leq P\}$.

Then, the proximity between node pair (s, t) and (x, y) by their correlation semantics is given as,

$$Rsim((x, y)|(s, t)) = 1/Euc(f_{(x,y)}, f_{(x,t)}) \quad (1)$$

where $Euc(a, b)$ is Euclidean distance between vector a and b. We will use the formula to achieve proximate node pairs searching in experiments and evaluate CMPM's effectiveness.

4 Combinational Meta-paths Mining

The implementation of Combinational Meta-Path Mining Algorithm (CMPM) is based on two basic assumptions: (1) shorter meta-paths generally carry more significant semantics, and longer meta-paths are good at ensuring semantic completeness; (2) path instance distribution in network is relevant to path importance, it can be used to weight meta-paths. Rational behind the first assumption is intuitive, long path can bring remote weak-connected neighboring nodes into path searching steps, which undermines the discriminative ability between node pairs [1]. For the second assumption, distribution of path instance is implicitly influenced by path length, using it as path weighting reference can balance semantics carried by meta-paths of different length. Follow the basic assumptions, we implement combinational meta-path mining algorithm in two steps: candidate meta-paths preliminary screening step and path refinement step, as showed by flow chart in Fig. 1.

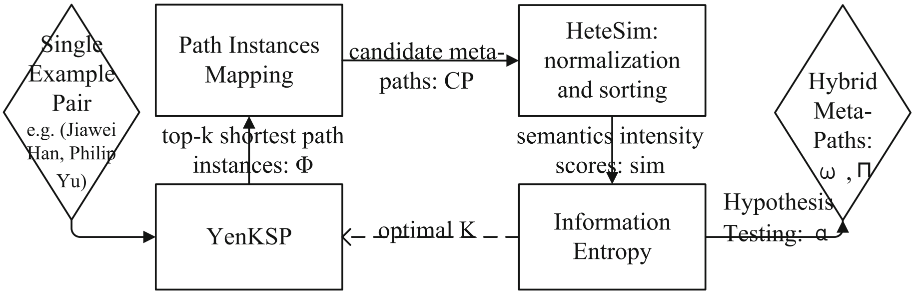

Fig. 1. Flow chart of CMPM implementation steps.

In candidate meta-paths preliminary screening step, CMPM uses YenKSP algorithm [16] to obtain representative path instances that connect nodes in example pair. YenKSP starts with shortest path algorithms to find path instance P^1 connects (s, t). Then, setting node in $P^{t-1}(2 \leq t \leq K)$ one after another as deviation point, and defining sub-path from source to deviation node as root path R_i^t. Subsequently, starting from deviation point to find another shortest path to target and define as spur path S_i^t. Finally, merge them as $P_i^t = R_i^t + S_i^t$ and push into a priority queue to pop the shortest one as P^t. Repeat the process until all top-k shortest path instances in Φ are found. We map paths in Φ to candidate meta-paths $CP=\{\Pi_{\mathrm{i}}, 1 \leq i \leq Q\}$, where Q is the number of candidate meta-paths. During the path mapping step, we also record number of path instances included with respect to each unique meta-path as $M = \{\Pi_1 : n_1, \Pi_2 : n_2, \ldots, \Pi_Q : n_Q\}$.

Path refinement is the main step of this paper. Number of path instances mapped to each meta-path is defined as $n_i = M[\Pi_i]$, it implicitly reflects significance of specific path by their degree of prevalence in HINs. We weight Π_i by $w_i = n_i / \sum_{i=1}^{Q} M[\Pi_i]$ to show relative meta-path importance. Now, weighted candidate meta-paths still contain redundancies. We can remove them by calculating their semantical significance. Using HeteSim, the semantic intensity score corresponds to $\omega_i \Pi_i$ is $\omega_i * \sigma(s, t|\Pi_i, G)$ as defined in Sect. 3, and we store them in vector sim. Next, scores in sim are normalized and sorted in descend order, and paths in CP are also updated accordingly to maintain one-by-one matching relationship between meta-path and corresponding semantic score. Semantic intensity scores in sim actually describe the likelihood that given example pair carries semantics specified by meta-paths in CP. This made it reasonable to use information entropy [17] to measure the expected semantics carried by combinational meta-paths. Define global entropy for candidate meta-paths CP as H, we use statistical hypothesis test to remove redundant meta-paths. Because each time we add meta-path from CP into combinational meta-paths, corresponding entropy will increase step by step until it covers $1-\alpha$ of H, where α (5%–10%) is the significance level. As a result, there exists P genuinely selected single meta-paths from *CP*, and we update their weights again using $\omega_i = n_i / \sum_{i=1}^{P} M[\Pi_i]$.

Algorithm 1: CMPM (G, s, t)

```
Input: network G, example pair (s, t)
Output: meta paths Π, weights ω
Φ←YenKSP(G, s, t, k);
foreach path instance φ in Φ:
  π←mapping φ to meta-path;
  M [π]←M [π] +1;
CP←candiate meta-paths from M;
foreach (π, n) in M:
  sim.insert(n/∑_{i=1}^{Q} M[Π_i]*σ((s,t)|Π,G));
sim=sim/∑_{i=1}^{Q} sim[i];
sim.sort();
CP←update candidate paths accordingly;
H = sim.entropy()
foreach similarity score s in sim:
  area←area-s*log(s);
  if area/H<1-α:
    Π.insert(CP[sim.index(s)]);
foreach π in Π:
  ω.insert(M[Π]/∑_{k∈Π} M[k]);
return Π, ω
```

According to description above, Algorithm 1 gives the detailed implementation of CMPM. The most time consuming part of CMPM is YenKSP, which occupies $O(KV(E+V\ \log V))$ if search for shortest path using Fibonacci heap [16], where K, E, V are path instance numbers, edge and node numbers. In path refinement step, the time complexity is $O(QL^2)$ if using dynamic programming based HeteSim [10] to measure semantic intensity scores, where Q is candidate meta-paths numbers and L is maximum meta-paths length among all meta-paths.

5 Experiments and Analysis

In this section, we perform proximate node pairs detection in HINs to validate the effectiveness of CMPM algorithm. In experiments, we use bibliographic networks ACM and DBLP from Artminer [14] archived in 2016. In preprocessing step, we remove data before 2006, and only reserve papers published in top conference related to data mining, information retrieval, database system and AI, finally only 16310 and 24332 papers are remained. In algorithm evaluation step, we use CMPM to mine author-similarity correlations and author-publication correlations between given example pairs. Equation (1) is used to detect proximate node pairs carrying similar correlation semantics as that defined by combinational meta-paths given previously. Because there has no benchmark for author-similarity correlation comparison, we collect paper-citations, h-index, i10-index and scholar-closeness data from Google Scholar as $\lambda = (citations, h\_index, i10\_index, ranking)$, and ranking benchmarks by calculating Euclidean distance with respect to $\lambda_{JiaweiHan}$,

$$\Delta = \|\lambda_{others} - \lambda_{JiaweiHan}\| \tag{2}$$

the smaller Δ is the researcher will be more similar with Jiawei Han.

In meta-paths mining step for author-similarity correlations, we choose (Jiawei Han, Philip S. Yu) as example pair, because they have high publication and citation index in data mining area, and their work also have been recognized by peers. Most importantly, they share no extra close correlation like teacher-student, colleges, classmates. Path mining results given by CMPM in our experiment shows that single meta-path A-P-T-P-A has the largest weight among combinational meta-paths, and A-P-A-P-A, A-P-C-P-A follows. This phenomenon suggests us that research interests, co-authorships and conferences are vital to describe researcher's similarity in academic capability.

Then, we detect proximate node pairs similar to (Jiawei Han, Philip S. Yu) using Eq. (1), and Jiawei Han is set as source node for simplicity. Nodes detection results constrained under CMPM based combinational meta-paths and prior knowledge based meta-paths (controlled groups have meta-paths APA, APTPA, APCPA and their simple combinations PriorComb) are listed in Table 1. It shows node pairs searching results under CMPM in both datasets are more likely to find top researchers such as Jian Pei, Xifeng Yan, Charu Aggarwal and Christos Faloutsos than that of controlled groups. This means, meta-paths given by CMPM do capture meaningful correlation semantics carried by example pair, and they also outperform prior knowledge based ones.

Table 1. Top 5 node pair proximity searching results for author-similarity correlation relationships with ACM and DBLP datasets.

Rank	CMPM	APA	APTPA	APCPA	PriorComb
	ACM datasets				
1	PhilipS.Yu	JiaweiHan	Jiawei Han	JiaweiHan	PhilipS.Yu
2	JianPei	XifengYan	PhilipS.Yu	PhilipS.Yu	Jia Pei
3	XifengYana	PhilipS.Yu	ChrisFaloutsos	MarianneWinslett	XifengYan
4	CharuAggarwal	JianPei	CharuAggarwal	ChrisFaloutsos	YizhouSun
5	ChrisFaloutsos	YizhouSun	JianPei	CharuAggarwal	ChiWang
	DBLP datasets				
1	PhilipS.Yu	JiaweiHan	PhilipS.Yu	JiaweiHan	PhilipS.Yu
2	CharuAggarwal	XifengYan	Jiawei Han	PhilipS.Yu	YizhouSun
3	JeffreyXuYu	YizhouSun	CharuAggarwal	CharuAggarwal	Dong Xin
4	ChristosFalouts	PhilipS.Yu	MarianneWinsle	JeffreyXuYu	XifengYan
5	XifengYan	Dong Xin	ChristosFalouts	ChristosFaloutsos	XiaoleiLi

Table 1 only shows a general profile of performance advance by CMPM, so we introduce normalized degree of disorder [18] as a measure,

$$D(\Phi, \Phi') = \begin{cases} 2/N^2 \sum_{k=t_1}^{t_n} |\Phi'[k] - \Phi[k]|, \ \text{n} \in \text{even} \\ 2/(N^2 - 1) \sum_{k=t_1}^{t_n} |\Phi'[k] - \Phi[k]|, \ \text{n} \in \text{odd} \end{cases} \tag{3}$$

where $\Phi = \{t_1 : 1, t_2 : 2, \ldots, t_n : n\}$ contains rankings in google benchmarks and Table 1.

It calculates normalized deviation between Φ and Φ', smaller deviation means better searching results. Precision, recall and F1-score are also used in experiments. Results under the four measuring schemas with ACM datasets are showed in Fig. 2. In (a), it shows degree of disorder for node pair searching results using meta-paths CMPM, APA and PrioriComb, where disorder value for CMPM constrained results consistently smaller than the prior knowledge based results, this means searching

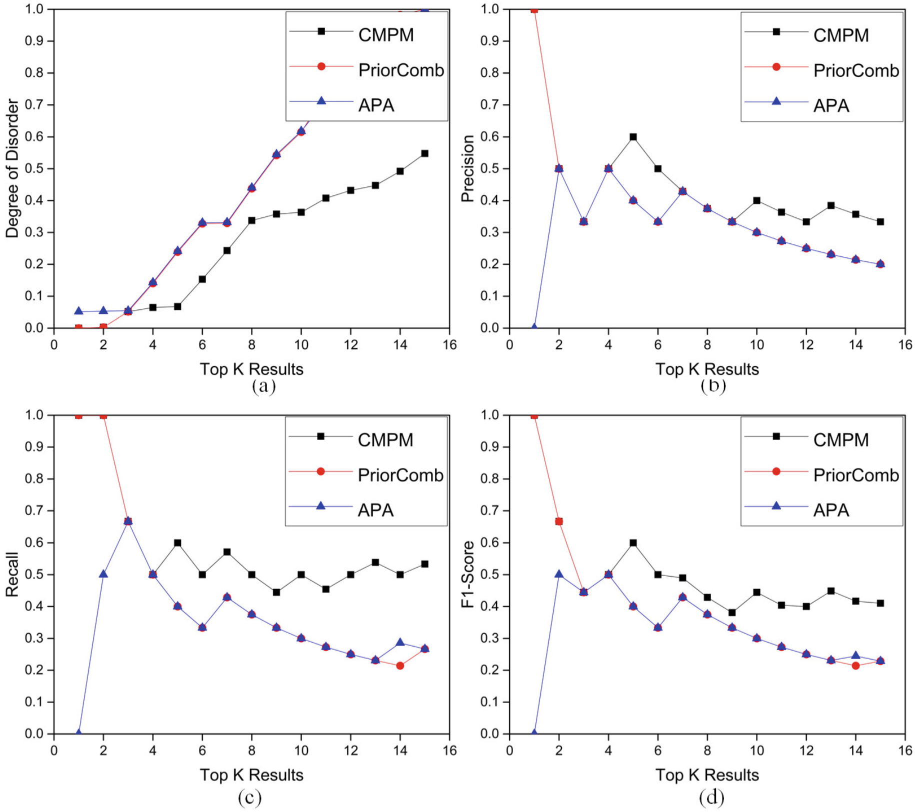

Fig. 2. Normalized degree of disorder, precision, recall, and F1-score measuring results for proximate node pairs detection results under ACM datasets.

results under CMPM will be closer to google benchmarks we proposed previously. Except for low degree of disorder using CMPM, we also find that the performance boost for CMPM will be enlarged when we request for more top K node pairs. This means, the overall searching results measured by degree of disorder will be better than that showed in (a). In (b) and (c), we use precision and recall as a measure. Results from precision and recall showed us that CMPM based searching results outperform both PriorComb and APA, and CMPM seems to have pretty good performance when measured with recall. To compromise the measures given by precision and recall, (d) is measured with F1-score, which prefers higher precision and recall. It's obvious to observe that CMPM has higher F1-score than the other two prior knowledge based meta-paths. Moreover, one common phenomenon under the four measures is that results for PriorComb and A-P-A are very close to each other, this signaling us that if meta-paths are not genuinely weighted, correlation semantics may at the risk of being covered by their their counterparts. On the other hand, top 3 searching results by PriorComb are close to CMPM based results, but their gaps gradually enlarged. This means prior knowledge based meta-paths do carry vital correlation semantics, but their depth is weaker than CMPM based combinational meta-paths. To validate scalability of CMPM, we use example pair (Jiawei Han, SIGKDD) to reveal author-publication correlation relationships. After similar experiment procedure, results are showed in Table 2. We rate the ranking result with scores between 0–2, represent irreverent, some-relevant and complete-relevant. Then nDCG [19] is used to measure ranking accuracy based on ranking positions in the list with scores between 0–1, the higher the better. It shows, in both datasets, CMPM still outperforms prior knowledge based PriorComb.

Table 2. Top 10 node pairs proximity searching results for author-publication correlation relationships, where nDCG scores are given below.

Rank	ACM		DBLP	
	CMPM	PriorComb	CMPM	PriorComb
1	SIGKDD	SIGKDD	SIGKDD	SIGKDD
2	SIGMOD	SIGMOD	TKDD	TKDD
3	VLDB	VLDB	TWEB	TWEB
4	SIGIR	SIGIR	DKE	EDBT
5	TKDD	TKDD	IPM	DKE
6	TODS	AI	EDBT	IPM
7	TOIS	CVPR	SIGMOD	SIGIR
8	TWEB	TODS	SIGIR	SIGMOD
9	SDM	ACL	VLDB	ICDE
10	WWW	TOIS	ICDE	WWW
nDCG	0.74	0.61	0.57	0.50

We also studied the influence of path instance numbers to YenKSP based preliminary screening, it shows entropy carried by candidate meta-paths increase steadily at beginning, but it slows down immediately when path instance numbers exceed some thresholds. This hints us to choose optimal K for YenKSP. All experiments are performed on Dell XPS 8900 with eight 3.4 GHz Intel i7-6700 CPUs and 16 GB of memory, the time consumed in CMPM based meta-paths mining for the author-similarity example pair are 132.9 s under ACM datasets and 153.1 s under DBLP datasets (initial path instance numbers are all set to 500).

In summary, author-similarity and author-publication based correlation relationships mining experiments in this section have proved the effectiveness of CMPM algorithm in combinational meta-paths mining. Experiment results showed us that meta-paths mined by CMPM carry more complete correlation semantics than prior knowledge based meta-paths, and it also captures subtler semantics. Additionally, path weighting schema in CMPM also balance semantics carried by each single meta-path, which prevents global semantics from being covered by single semantically significant meta-path.

6 Conclusion

In this paper, we propose a new algorithm CMPM to mine combinational meta-paths that describe correlation relationships between nodes in HINs. The algorithm is based on two assumptions: (1) shorter meta-paths generally carry more significant semantics, but longer meta-path is helpful to guarantee semantic completeness; (2) path instance distributions in networks is relevant to path importance, we can use them to weight meta-paths. Then a preliminary screening step and a path refinement step are included in its implementation. In experiment, author-similarity and author-publication correlation relationships in bibliographic networks are evaluated. Corresponding nodes proximity detection results measured by average degree of disorder, accuracy, recall, F1-score and nDCG proved the effectiveness of CMPM algorithm. In the future, we will further improve the algorithm by propose better path weighting methods, and extend the algorithm to other HINs also deserve efforts.

Acknowledgements. This work is supported by National Key Research and Development Plan in China (2017YFB0803300), National Natural Science Foundation of China (U1636105).

References

1. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: meta path-based top-k similarity search in heterogeneous information networks. Proc. VLDB Endow. **4**, 992–1003 (2011)
2. Cao, B., Kong, X., Philip, S.Y.: Collective prediction of multiple types of links in heterogeneous information networks. In: 2014 IEEE International Conference on Data Mining, pp. 50–59. IEEE (2014)
3. Cao, X., Zheng, Y., Shi, C., Li, J., Wu, B.: Meta-path-based link prediction in schema-rich heterogeneous information network. Int. J. Data Sci. Anal. **3**(4), 285–296 (2017)

4. Vahedian, F., Burke, R., Mobasher, B.: Meta-path selection for extended multi-relational matrix factorization. In: The Twenty-Ninth International Flairs Conference (2016)
5. Shi, C., Li, Y., Philip, S.Y., Wu, B.: Constrained-meta-path-based ranking in heterogeneous information network. Knowl. Inf. Syst. **49**(2), 719–747 (2016)
6. Vahedian, F., Burke, R., Mobasher, B.: Weighted meta-path generation, multi-relational recommender system, heterogeneous information network, weighted random walk sampling (2017). ArXiv preprint arXiv:1703.00034
7. Liang, T., Chen, L., Wu, J., Dong, H., Bouguettaya, A.: Meta-path based service recommendation in heterogeneous information networks. In: International Conference on Service-Oriented Computing, pp. 371–386. Springer (2016)
8. Shi, C., Zhou, C., Kong, X., Yu, P.S., Liu, G., Wang, B.: Heterecom: a semantic-based recommendation system in heterogeneous networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1552–1555. ACM (2012)
9. Gupta, M., Kumar, P., Bhasker, B.: HeteClass: a meta-path based framework for transductive classification of objects in heterogeneous information networks. Expert Syst. Appl. **68**, 106–122 (2017)
10. Shi, C., Kong, X., Huang, Y., Philip, S.Y., Wu, B.: Hetesim: a general framework for relevance measure in heterogeneous networks. IEEE Trans. Knowl. Data Eng. **26**, 2479–2492 (2014)
11. Lao, N., Cohen, W.W.: Relational retrieval using a combination of path-constrained random walks. Mach. Learn. **81**, 53–67 (2010)
12. Meng, C., Cheng, R., Maniu, S., Senellart, P., Zhang, W.: Discovering meta-paths in large heterogeneous information networks. In: Proceedings of the 24th International Conference on World Wide Web, pp. 754–764. ACM (2015)
13. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 413–422. ACM (2013)
14. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 990–998. ACM (2008)
15. Zhang, J., Yu, P.S., Zhou, Z.-H.: Meta-path based multi-network collective link prediction. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1286–1295. ACM (2014)
16. Yen, J.Y.: Finding the k shortest loopless paths in a network. Manag. Sci. **17**, 712–716 (1971)
17. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Comput. Commun. Rev. **5**, 3–55 (2001)
18. Gunetti, D., Picardi, C.: Keystroke analysis of free text. ACM Trans. Inf. Syst. Secur. (TISSEC) **8**, 312–347 (2005)
19. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. (TOIS) **20**, 422–446 (2002)